How People Judge the Usability of a Desktop Graphic User Interface at Different Time Points: Is there Evidence for Memory Decay, Recall Bias or Temporal Bias?

How people judge the usability of a desktop graphic user interface at different time points: Is there evidence for memory decay, recall bias or temporal bias?

Kyle Boyd*
Ulster University
Belfast, BT15 1ED
+44 28 9536 7205
ka.boyd@ulster.ac.uk

Raymond Bond
Ulster University
Jordanstown, BT37 0QB
+44 28 9036 8156
rb.bond@ulster.ac.uk

Attila Vertesi
Bournemouth
& Poole College
Poole, BH14 0LS
+44 1202 205313
vertesia@bpc.ac.uk

Huseyin Dogan
Bournemouth University
Dorset, BH12 5BB
+44 1202 962491
hdogan@bournemouth.ac.uk

Justin Magee
Ulster University
Belfast, BT15 1ED
+44 28 9536 7015
jdm.magee@ulster.ac.uk

Research Highlights

- The system usability scale is used to test for memory decay and temporal bias in judging the user experience of technologies at different time points.
- 212 participants took part in two studies ranging from 3 weeks to 6 months.
- There is no evidence that there is a temporal bias or memory decay when users complete a SuS survey at the two different time points of 3 weeks and 6 months.

Abstract

The System Usability Scale (SuS) survey is a widely respected tool for measuring usability. Generally, a SuS score is administered directly after a usability test to assess the usability and user experience of digital products. However, some researchers have used SuS as a survey as part of longitudinal 'in the wild' trials where SuS is often completed some period after the trial. The aim of this research was to determine if a participant's memory of a product's usability would change if a SuS survey was administered at different times after a test. Hence, we sought to understand if recalling the usability of a digital technology was affected by temporal bias or memory decay. This paper includes results and findings from two studies, study 1 involved evaluating a web application and study 2 involved evaluating a virtual learning environment. Collectively the two studies had 212 participants (n=212). The findings conclude that there is no significant change of the user's recollection of the usability of digital product as evidenced by an analysis of users who completed multiple SuS surveys over a short term period of 3 weeks or over an extended period of time of 6 months.

Usability, System Usability Scale, User Experience, Usability Testing, Human-Computer Interaction, User Interfaces

1. INTRODUCTION

User Experience (UX) as a discipline has evolved considerably over the last number of decades. The introduction of mediums such as mobile and the web including, native, audio and tactile input means that over time the process of how we conduct UX design has changed. As a discipline, designers design experiences and the aim is to make these experiences better (Murphy 2018). The UX design process is an iterative process of Observation → Idea Generation → Prototyping → Testing. This loop continues through multiple iterations to ensure that the designer's assumptions are tested and possible solutions are developed. By trying to understand users better, there has been a drive toward UX research via the testing phase, which involves measuring usability. Usability is a key sub-construct of UX and refers to the process of evaluating a digital product or service by testing it with representative users (Nielsen and Norman 2012, Boyd *et al.* 2017, Bond *et al.* 2014).

There has been some ambivalence regarding how the terms 'UX' and 'usability' are related. The Usability ISO 9241 definition states that usability is "the effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments" and the definition of UX in ISO FDIS 9241-210 states that "UX includes all the users emotions, beliefs, preferences, perceptions, physical and psychological responses that occur before, during or after use".

Vermeeran et al. analyses the different UX evaluation methods but also discusses the differences between UX and usability (Vermeeran 2010). They state that these two constructs are intertwined. Usability is subsumed by UX implying that UX evaluation must go beyond the existing methods of

usability evaluation. Usability testing normally focuses on ease-of-use and task performance but these are not sufficient for UX. In UX evaluation, it is important to determine how the user feels about the system. Whilst the user satisfaction component of usability testing can be seen as UX, there are other areas such as motivations and expectations which usability testing does not measure. In this paper, we take the stance that usability has been subsumed within UX (Bevan 2009, Norman, 2009). In reality, UX is an elaboration of the satisfaction component of usability and is an umbrella term for the user's perceptions and responses whether measured subjectively or objectively. Regardless of the terminology used, UX and usability have two objectives which are to optimise human performance and user satisfaction (Bevan 2009). Whilst this study utilises usability methods, it can be accepted that enhanced usability will create a better UX.

There are a broad range of methods to evaluate the usability of a system (Curendale 2013) such as expert heuristic evaluations, cognitive walkthroughs, benchmark testing and eye tracking analysis to name but a few. However, the think-aloud usability test has become a popular method. During a think-aloud usability test, participants will try to complete tasks while observers watch, listen and take notes. The goal of the test is to identify usability problems, collect qualitative and quantitative data and determine the participant's satisfaction with the technology. To execute an effective usability test, it is necessary to develop a repeatable test protocol including user tasks/scenarios, appropriate participant recruitment and analytical reporting. Usability testing, often using incomplete or sketch prototypes, permits a process where proposed designs or individual features within a system are forced to fail early, fast and often, in order to refine the most robust user experience or effective functionality. This agile process fosters design thinking whilst identifying problems before a full product is designed and released for end use (Vermeeran 2010, Bevan 2009). The earlier the usability issues are identified, the sooner they can be rectified, resulting in less impact on time and cost. Typically, a usability test will assess:

- Effectiveness (the extent to which users can complete their tasks and achieve their goals successfully)
- Efficiency (the extent to which they expend resource in achieving their goals)
- Satisfaction (the level of comfort and/or enjoyment of the experience in achieving those goals)

It is important to collect the right data for affective analysis allowing for re-designs and recommendations to be made. There are various methods to derive the above attributes but one of the most popular and widely used methods are usability surveys such as the System Usability Scale (SuS).

1.1 SYSTEM USABILITY SCALE

SuS was created by John Brooke (Brooke 2013) in 1986 and allows for the evaluation of hardware, software, mobile devices, websites and general digital applications. SuS consists of a 10-item questionnaire, each presenting a Likert scale (normally 5 points) ranging from strongly agree to strongly disagree. Subsequently, a usability SuS score is computed. The standard SuS consists of the following ten items (odd numbered items are worded positively and even numbered items are worded negatively). Questions are as follows:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before i could get going with this system.

Typically, a SuS questionnaire is given to a user after they have completed a usability test allowing them to rate the usability of the technology (Krug 2009). However, researchers have been using this questionnaire in various ways, for example they have used SuS after a longitudinal study involving a trial of technology or directly after a lab-based usability test that have specified tasks or even after a session without tasks where a user casually reviews an app or a digital technology. These variations use different time points as to when the SuS is administered. In their guidelines entitled 'Applying Human Factors and Usability Engineering to Medical Devices', the US Food and Drug administration (FDA) state that memory decays over time, therefore any information collected at a particular time or point may not be as accurate or complete as it could be (Med Device Online 2018). It is suggested that if training is needed to train users on the use of a medical device, then a 'decay gap' between the training and the usability testing should be instituted. A decay gap could be an hour or even days. This ensures that the usability results are not biased by the training but reflects real world usage. The rationale for this comes from the 'Ebbinghaus forgetting curve' (Murre and Dros 2015), which suggests that humans forget fast soon after a stimulus but then diminish their 'forgetfulness' at a slower rate as time passes. However, given a UX could be an emotional experience, perhaps this type of experience is much different when compared to an ordeal whereby one must use knowledge and facts. According to Angelou "people may not remember exactly what you did, or what you said, but they will always remember how you made them feel" (Quote Investigator 2018). Moreover, given Norman (Norman 2004) details that the UX functions on three phases: 1) visceral, 2) behavioural and 3) reflective, perhaps the reflective phase can affect the judgement of a past user experience. The reflective aspect is when a user consciously weighs up the pros and cons of a technology. Reflecting on what it means for them to use the technology and whether they will continue to do so.

Temporal bias involves the change in human opinion over time, for example, a common temporal bias is hindsight bias (Sanna and Schwartz 2004) which may affect the user's opinion about the usability of a system on hindsight. Recall bias is well known in medicine (Sedgwick 2014) where patients or relatives of patients exaggerate past circumstances, symptoms and other important details about a past physical experience. Indeed there is also a well-known bias when recalling eye witness testimonies in the judicial system where witnesses are known to report details that were non-existent (Buckhout 1974). We hypothesise that these types of phenomena may affect SUS scores that are provided long after the UX, trial or usability test.

2. PREVIOUS WORK VALIDITY AND RELIABILITY

Bangor *et al.* (Bangor *et al.* 2008) conducted usability studies on various products and services using the SuS score. They conducted over 200 studies with 2300 surveys and found that the mean SuS score was 70 and the median was 75. Bangor *et al.* (Bangor *et al.* 2009) also analysed the interpretation of SuS and added new descriptors. They compared the SuS score with perceived levels of usability. A score over 85 was classified as excellent, a score of 70-85 was classified good to excellent, a score of 50-70 is acceptable but encompasses issues that need addressed and a score below 50 is classified as unusable and unacceptable. Tulis and Stetson (Tulis and Stetson 2004) measured the usability of two websites using a range of different surveys including the Questionnaire for User Interaction Satisfaction (QUIS), SuS and the Computer System Usability Questionnaire (CSUQ). It was found that the SuS provided the most reliable results across a range of samples.

Given the authors are interested in the user's recollection and their reflection of a past UX, we have identified a number of relevant studies. Koon *et al*. (Koon *et al*. 2013) explored the utilitarian, hedonic and social aspects of smartphones to measure how users continually engage in smartphone activity. There is also a body of work regarding how UX is affected over time. Norman argues that these memories of experiences will be reported to others and guide the future behaviour of the individual. Thus reconstructed memories are relevant despite potential bias (Norman 2009). Karapanos *et al*. conducted a five week study which followed six individuals during an actual purchase of an Apple iPhone (Karapanos *et al* 2009). The study found that prolonged use was motivated by different qualities than the ones provided initial positive response. While hedonic experiences were key early

on, the prolonged experiences became increasingly tied to how the product became meaningful in one's own life. Moellendorf *et al.* considered mobile phone usage over the period of more than a year (Moellendorf *et al* 2006). It revealed characteristic changes. Pragmatic utility perceptions remained stable and the usability of the product increased. Whereas hedonic perceptions (stimulation, beauty) deteriorated. Stimulation showed an increase in deterioration because of increasing habitation but beauty and identity induced a comparison with other newer products owned by the users.

Karapanos *et al.* presented iScale for retrospective elicitation of longitudinal UX studies (Karapanos *et al.* 2012) It was motivated by how people reconstruct emotional experiences from memory. They found that there was an increase in the richness and consistency of recalled information compared to recall. This provides support around the viability of retrospective techniques compared to longitudinal studies. They suggest that people communicate and act on their own biased memory and not on an unbiased objective summary of what happened. In supporting design, arguably, it maybe more important to understand what users remember rather than what was experienced. Oishi and Sillivan show that retrospective evaluations predict human behavior later on (Oishi and Sullivan 2006).

Further work by Kujala *et al.* called UX Curve was an exploratory study which was developed as an easy to apply method for supporting users in recalling important details of product qualities that affect the UX (Kujala *et al.* 2011). UX Curve relies on the memories of the user experiences which are retrospective. They found that users became more attracted to their mobile phones showing that UX can improve over long time periods.

We wish to further this body of work and differentiate by studying the usability of a technology over time using the SuS survey [McLellan *et al.* 2012 and Folstad 2017). User opinions in the moment of using a technology and retrospectively are likely to be different which is why some researchers prefer to use ecological momentary assessments (Moskowitch and Young 2006). In order to determine if a user's recollection and memory of a UX changes over time, a suitable protocol for repeatable usability testing was developed.

This paper highlights two studies that carry out multiple SuS analyses over a short time period of three weeks (Study 1) and a long period of six months (Study 2) to determine if SuS scores change over time. The paper includes an analysis of 212 participants. Study 1 analyses the recall accuracy of a usability study, retrospectively after 3 weeks have past. Study 2 is slightly different and analyses changes in the user's judgement of the usability of a technology over a 6 month period, hence comparing initial SuS scores at 0 months and at 6 months.

It is the authors' intentions to explore whether the user's memory of the usability changes over time using SuS scores as a metric. Our assumptions are that SuS scores at different time points can be impacted by memory decay or recall bias (Study 1), and that the user's opinion can change for the better or for the worse when repeatedly using the same technology over a longer time period (Study 2). If the user's opinion does not change over a longer time period (Study 2), then a lab-based usability test is a cost-effective approach perhaps removing the need for longitudinal studies (Karapnos *et al.* 2009).

In Study 1 the researchers measured the usability of a web application and invited participants to complete a SuS score immediately after the test and then over the following two weeks. The latter two time points involved the user completing the SuS survey using their memory of their past user experience (Boyd *et al.* 2018).

Study 2 focuses on using SuS to measure usability of a Virtual Learning Environment (VLE) after initial use in a lab and after having used it for up to 6 months (Vertesi *et al.* 2018). The list of tasks completed by participants was captured to identify the level of involvement for the post-study (i.e. after 6 months). Some users completed an online SuS survey (n=170) and others completed a paper-based copy (n=13).

2.1. RESEARCH QUESTIONS

The research questions are as follows:

(I) Does the memory and recollection of the usability of a technology change over time?

(II) Can users accurately recollect a past UX of a technology when three weeks have past? Users use the technology once and then report on their past experience at three time points.

(III) Do users change their judgement of the usability of a technology after having used it for up to 6 months?

3. METHODOLOGIES

The following section outlines the methodology of Study 1 and Study 2 with details of participants and data analysis. Study 1 was approved by the ethical approval by the Art & Design Research Ethics Committee (Ulster University) on 28th February 2018 and Study 2 was approved by the University Research Ethics Committee (UREC) of Bournemouth University on 25th May 2018.

3.1 STUDY 1: DATA COLLECTION

Participants were asked to complete a series of tasks (See Table 1) on a Web Application called Virtuagym (http://www.virtuagym.com) a publicly available web application which promotes healthy living (See Figure 1). It was our intention to have rudimentary tasks and that was perceived to have a neutral emotive experience. This was to focus participants to determine design inconsistencies and usability problem areas within the user interface and content areas.

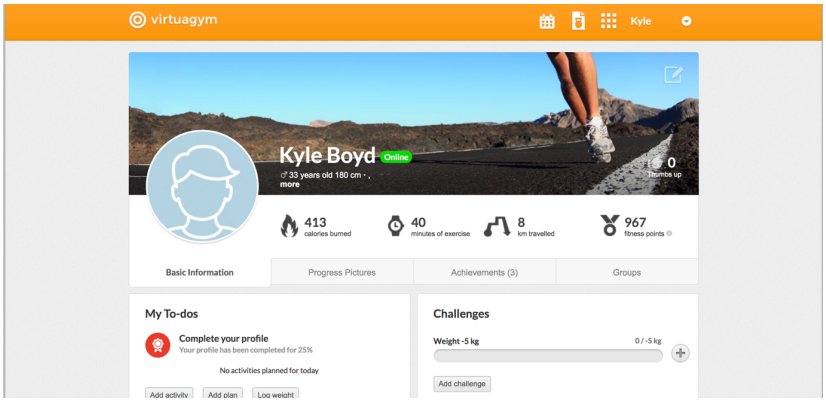

Figure 1: *The Website Virtuagym.com which was used in the study*

After each participant completed the tasks, they completed a SuS survey. Participants were sent another SuS survey via email at one week and two weeks after the usability test. When all the SuS questionnaires were completed, the data was collated and analysed using R Studio.

Table 1: *The tasks that were completed by participants using virtuagym*

Task Number	Task to Complete
1	Sign up to http://www.virtuagym.com
2	Go through the setup process
3	Add activity calendar and workouts to your portfolio
4	You want to go running each Saturday add a running activity
5	Each Tuesday and Thursday you go to the gym add a gym workout to the calendar

6	You would like to tone arms for the summer. Include a dumbbell weekly workout
7	You would like to raise money for charity – you are going to do 150 sit-ups a day. Add this challenge to your workout
8	Find out how many calories will be burned with this exercise regime?

13.2 STUDY 1: PARTICIPANTS

Study 1 recruited from a convenience sample totalling thirty participants from Ulster University who were invited to undertake the usability test. Participants were recruited from the BDes (Hons) Interaction Design course at the Belfast School of Art, hence there will be a context of IT proficiency bias in this group. The test took place in public buildings in Northern Ireland. Public buildings are chosen specifically as they are required by law to be accessible for those with disabilities ensuring participant inclusivity (Hepple 2010). This was an evaluative study and therefore no statistical analysis was used to model participant sample size. Within usability testing, sample sizes of between 5 and 15 are deemed appropriate, with 5 subjects yielding 80% of the usability issues (Nielsen 2003). The participants were given an information sheet and a consent form to provide them with an opportunity to review the study and ask any questions before the test. Written informed consent was obtained before commencing the study.

The study was conducted with 18 male and 12 female participants. Of those, one was aged between 25-34 and the remaining subjects were aged 18-24. When the participants were asked to self-evaluate their computer literacy (1 being novice and 5 being expert), 83% responded between 4 and 5. A total of 50% of the participants felt that learning a new technology was easy and 93% of user had frequently used technologies such as smartphones and tablets. Of the thirty participants, 63% felt that digital technology was important to accomplish tasks of daily living.

3.3 STUDY 1: RESULTS

There were three points of data collection through survey submissions to monitor cognitive performance. A total of 76 SuS survey completions were collected. This comprises of 33 SuS survey completions at time point one (immediately after the test), 25 completions at time point two (one week after the test) and 18 completions at time point three (two weeks after the test). There was 24% drop out after week 1 and 45% dropout after week 2. SuS distributions at time point two and three are not normally distributed (Shapiro test, $p<0.05$) whilst the SuS distribution at time point 1 maybe normally distributed ($p=0.1141$) perhaps due to sample size. Figure 2 shows that the median SuS score remained similar across all three time points. Median scores did increase slightly (22.50, 25, and 23.75). However, a Wilcoxon signed rank test showed that there was no statistical significance ($p>0.05$) between the three SuS distributions at the three time points (all p-values were above 0.3). Interquartile Ranges (IQRs) across three time points are 22.5, 15 and 14.375 respectively. In agreement with the median, the mean SuS scores slightly increased from time point 1 to time points two and three (mean SuS scores were 24.92, 26 and 25.13 respectively). However, there is no statistical significance between these distributions and the subtle change would have no inferential changes regarding usability classification, i.e. all average SuS scores across all three time points yield the same interpretation regarding the usability of the system.

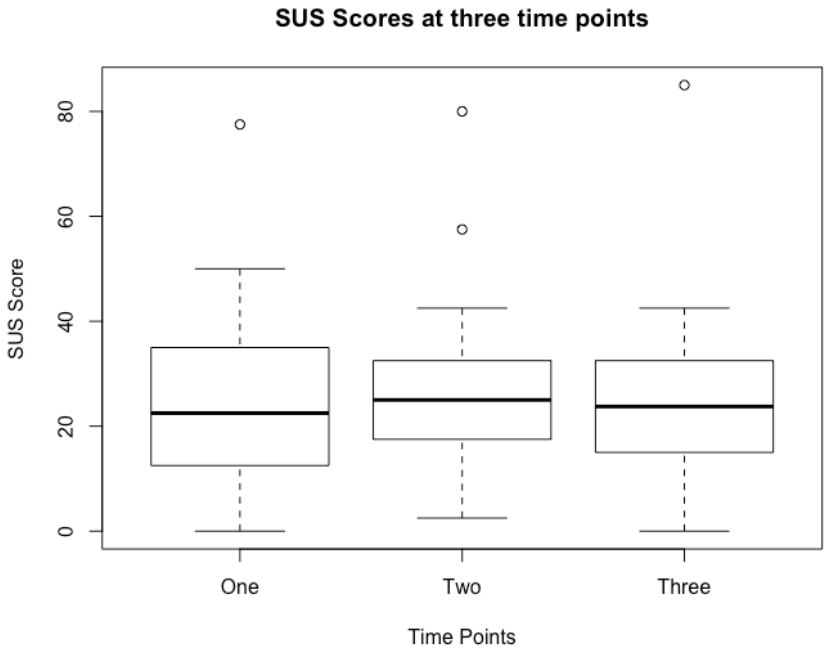

Figure 2: *Boxplots of SuS scores across three time points.*

Standard deviation of SuS scores across time points are as follows: 16.81, 17.31, 19.24 respectively. This shows a slight increase in variance as time progresses. Levene's test for homogeneity of variance indicated a statistically significant difference between the variance at time point 1 and the variance at time point 3 ($p<0.001$). However, whilst the variance is different, this is perhaps due to outliers and this change in variance would not be sufficient in effecting the interpretation of the system's usability based on SuS scores and current SuS benchmarking.

Figure 3 shows the boxplots of each SuS question at each test time. Questions 2, 8, 9 and 10 seem to have different medians at the different time points.

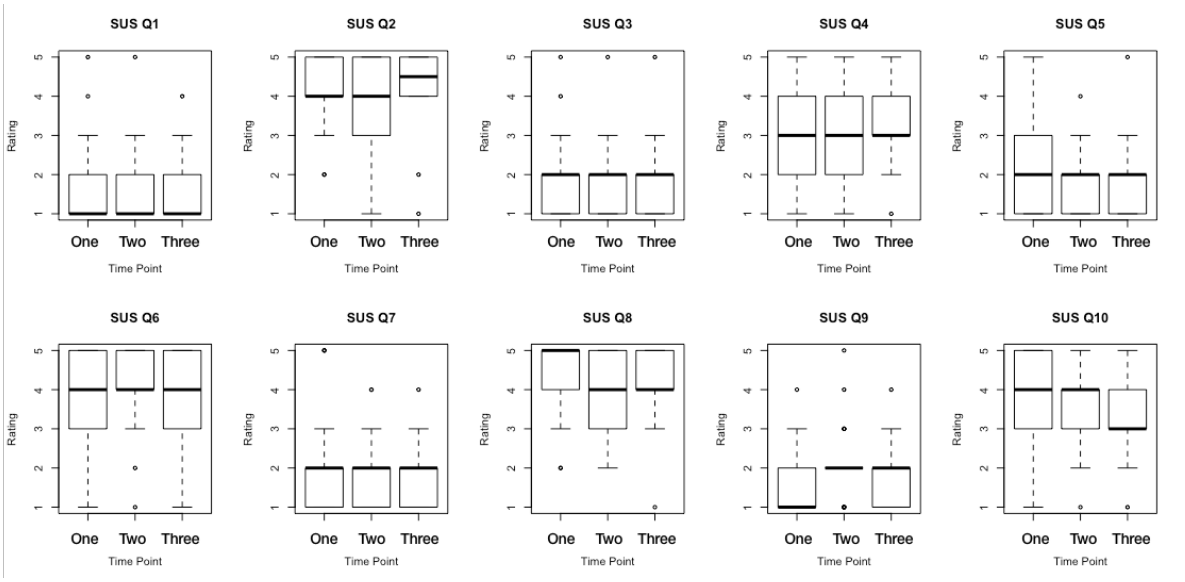

Figure 3: *The Boxplots showing ratings for each SuS question at each time point.*

This work suggests that retrospective SuS scores regardless of when administered after a test can provide similar results (up to at least 3 weeks later). This goes someway to validate recalling impactful

user experiences with a web application. These memories are more important than the actual experiences and affect the participants' attitude to the product (Karapnos *et al.* 2009). This is reasonable to suggest with study 1. Participants reported that the hedonic experience of the web application was poor as was the utility as reported by the SuS.

4.1 STUDY 2: DATA COLLECTION

The specific user groups (students, administrators, academics and learning technologists) performed the tasks based on the most common activities they needed to accomplish using the Virtual Learning Environment (VLE). Table 2 shows the task lists for the user test.

Table 2: *The tasks that were completed by students, administrators and academic/learning technologists*

Task Number	Student Tasks	Administrator Tasks	Academics and Learning Technologists
1	Access a unit area within the VLE	Navigate to a unit area	Take three word documents and make them available to students.
2	Review unit announcements for any notices	Take three word documents and make them available to students	Make some text and an image available to students.
3	View on line unit material available within the unit	Make a document unavailable to students	Create a link to an external website and make it available to students.
4	Open word documents made available	Create a link to an external website and make it available to students	Make a YouTube video available to students.
5	View embedded/linked video content	Post an announcement to students enrolled on the unit	Edit one of the items created in steps 1-4.
6	View the unit discussion topic and post an introductory message	Send an email to the students enrolled on the unit	Re-organise the items previously created.
7	View the unit blog and post an introductory post	Create a group of students for the unit	Make one of the items created in steps 1-4 unavailable to students.
8	View the unit wiki and post an introductory page	View student grades and assessments	Post an announcement to students.
9	Complete the sample unit test	Access an individual Turnitin submission, view grade and feedback.	Send an email to the students enrolled on the unit.
10	Submit an assignment via Turnitin	Add a grade for a non-Turnitin student assessment	Create a group of students for the unit.
11	View your grades	Add grades for all students on a non-Turnitin student assessment	Create a discussion topic and post an introductory message.
12	View any notifications	Use the grading functionality to create a calculation which sums the Turnitin and non-Turnitin assessments	Create a blog and post an introductory post.
13			Create a wiki and post an introductory page.
14			Create a test containing one multiple choice question and one multiple answer question.
15			View student grades and assessments.

16			Access an individual Turnitin submission, add a grade and feedback.
17			Add a grade for a non-Turnitin student assessment.
18			Add grades for all students on a non-Turnitin student assessment.
19			Use the grading functionality to create a calculation which sums the Turnitin and non-Turnitin assessments.

The pre-study was undertaken in a controlled lab environment when the VLE was first introduced. The participants were given some time (approx. 10 minutes) to familiarise themselves with the VLE. The data was collected through both an online and paper-based survey utilising the standard SuS questions. The post-study was then administrated after 6 months of the introduction of the VLE and involved the same set of tasks and procedure for data collection. A tick-list of tasks was also captured to identify that the tasks had been completed during the 6 months.

4.2 STUDY 2: PARTICIPANTS

Staff and students were invited to take part in the study. Participation was on a voluntary basis and participants details were kept anonymous. The pre-study involved 81 participants: students (n=40), academics (n=32), learning technologists (n=5) and administrators (n=4). The post-study which took place after 6 months involved 182 participants: students (n=137), academics (n=23), learning technologists (n=3) and administrators (n=19). Printed (paper) and online questionnaires were offered.

4.3 STUDY 2: RESULTS

Figure 4 shows that there is a difference in ratings for questions 1 ($p<0.01$), 4 ($p<0.01$), 6 ($p=0.14$) and 9 ($p=0.01$) when comparing ratings from users with no longitudinal experience with the VLE system, verses users who had 6 months experience with the system. Figure 5 shows the differences in SuS scores from users with and without longitudinal experience of the system and indicates that there is no statistically significant differential ($p=0.338$). However, there does seem to be a slight increase in SuS scores for the users who had 6 months experience in using the VLE system but this is not statistically significant. This work suggests that lab-based usability testing and the use of SuS with and without users who have had longitudinal experience can provide similar results. These findings validate the time restricted lab-based usability testing given it provides similar SuS scores as ascertained from a longitudinal usability study.

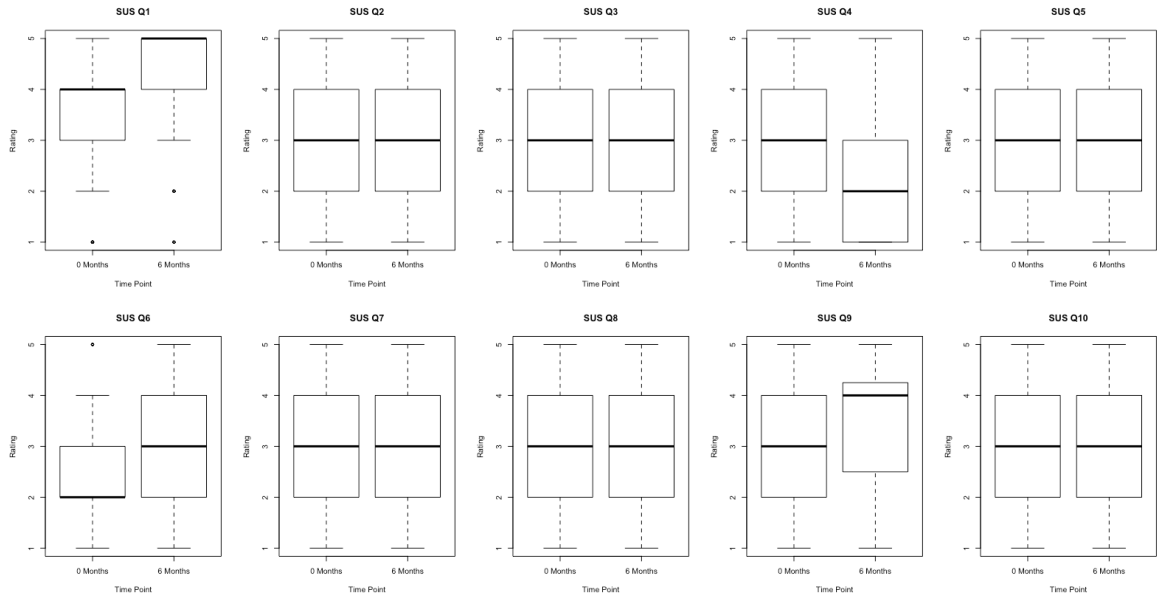

Figure 4. *Ratings for each SuS question from users who had no longitudinal experience with the system (VLE) and ratings from users who had 6 months experience of using the system.*

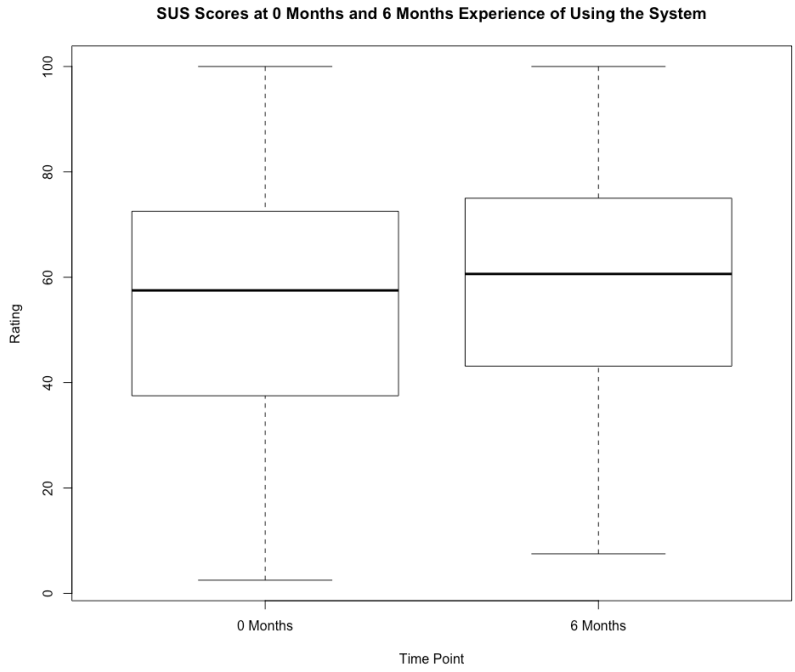

Figure 5*. SuS scores from users with and without 6 months experience of using the system.*

5. DISCUSSION

The authors intended to understand the recall of users judging the usability of a system using SuS. This paper presents two studies which aimed to consider if a UX judgement can change over time. The first study involved participants who used a web app once and then reported their user experience over a three week period post test. The second study involved an educational VLE over a time period of six months where users reported the usability at 0 months and at 6 months. Both studies used the SuS survey to report their findings.

Firstly it is necessary to address the methodology of using a usability tool such as SuS to measure a user experience. The reason for this is that the authors purport that UX is subsumed by usability which is supported by others (Bevan 2009). Usability focusses on usability issues, the completion of tasks, time taken and the user satisfaction. UX is also arguably an expansion of the satisfaction component of usability.

Using G* Power software, we have carried out an apriori statistical power calculation to inform future research (based on the difference of two dependent means [the mean SUS scores]). To achieve statistical power of 95% when expecting an effect of 0.5 (a projected moderate difference between mean SUS score at baseline and mean SUS score 2 weeks after using the web application), we would require a sample size of 54, which was not achieved in Study 1. This is a clear limitation of this paper. However, to achieve statistical power of 75%, a sample size of 30 is required which is a similar sample size achieved in Study 1. Nevertheless, the percentage difference between mean SUS score at baseline and at 2 weeks in Study 1 is a mere 0.85% (24.92 vs. 25.1). This indicates that if there is a change in SUS score, it is very small and insignificant in terms usability grading.

From the two studies completed, the authors found that, statistically, there was no memory decay or recall bias in recollecting the usability of a technology over a 3 week period. We also found that there was no form of temporal bias on the user's usability judgements over a 6 month period even when users continued to use the product. This could be due to the nature of the software being used in these studies and the complexity of the tasks used, therefore yielding no changes in usability grading. The two studies show that retrospective usability tests, of up to a period of 6 months using SuS, can act as an alternative to longitudinal studies but garner similar findings as supported by other works by Karapanos *et al.* (Karapanos *et al* 2012).

To build up a body of work to inform the expert's choices of which usability tool to use for which particular tests (Kortum and Sorber 2015, Orfanou *et al.* 2015 and Lewis 2014.), future work includes stress testing the SuS survey even further by answering the following questions:

(I) Task orientation: Is there a variation in the memorability of SuS scores when comparing a structured schedule of user tasks against casual browse and retrieval methods?

(II) Is there a variation in SuS scores when using different usability questionnaires for the same task? We would also like to conduct the same test with the range of usability questionnaires.

(III) Considering emotional design factors (Desmet & Hekkert), does fun, enjoyable and desirable user interfaces result in improved memorability?

(IV) Does age and/or IT proficiency effect the recall of a past UX, due to increased cognitive load during completion of the user test?

6. CONCLUSION

There is no evidence that there is a temporal bias, recall bias or memory decay when users judge the usability of a technology, at least over a short period of time (3 weeks) or over an extended period of time (6 months) (Kujala *et al.* 2011). Hence practitioners should not be overly concerned about the time at which subjects complete the SuS survey. However, limitations include the fact that there was subject drop out across the last two time points in Study 1. We also did not achieve a statistical power of 95% given that there were limited participant numbers. Some insignificant findings include the fact that SuS scores increased very slightly over time along with the variance of SuS scores, however this change would not alter usability grading using SuS.

7. REFERENCES

Bangor, A., Kortum, P., and Miller, J. (2008). An empirical evaluation of the System Usability Scale. International Journal of Human-Computer Interaction, 24(6), pp574–594. doi:10.1080/10447310802205776

Bangor, A., Kortum, P., and Miller, J. (2009). Determining what individual SuS Scores mean: Adding an adjective rating scale. Journal of Usability Studies, 4(3), pp114–123.

Bevan, N., 2009, August. What is the difference between the purpose of usability and user experience evaluation methods. In Proceedings of the Workshop UXEM (Vol. 9, pp. 1-4).

Bond, R.R., Finlay, D.D., Nugent, C.D., Moore, G. and Guldenring, D. (2014). A usability evaluation of medical software at an expert conference setting. *Computer methods and programs in biomedicine*, *113*(1), pp.383-395. doi: 10.1016/j.cmpb.2013.10.006.

Boyd, K., Bond, R. R., Magee, J., & Mc Cormack, P. (2018). Can users recall their user experience with a technology? Temporal bias and the system usability scale. In Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI 2018) http://dx.doi.org/10.14236/ewic/HCI2018.25

Boyd, K., Bond, R., Gallagher, S., Moore, G., & O'Kane, E. (2017, July). Usability and behaviour analysis of prisoners using an interactive technology to manage daily living. In *Proceedings of the 31st British Computer Society Human Computer Interaction Conference* (p. 80). doi: http://dx.doi.org/10.14236/ewic/HCI2017.80

Brooke, J. (2013) "SuS : A Retrospective," J. Usability Stud., vol. 8, no. 2, pp. 29–40. DOI?

Buckhout, R. (1974). Eyewitness Testimony. Scientific American, 231(6), 23-31. Retrieved from http://www.jstor.org/stable/24950236

Curedale, R. A. (2013). *Interviews Observation and Focus Groups: 110 Methods for User-centered Design*. Design Community College Incorporated.

Følstad, A., (2017). Users' design feedback in usability evaluation: a literature review. *Human-centric Computing and Information Sciences*, *7*(1), p.19. https://doi.org/10.1186/s13673-017-0100-y

Hepple, B. (2010). The new single equality act in Britain. *The Equal Rights Review*, *5*, pp.11-24.

Karapanos, E., Martens, J. B., & Hassenzahl, M. (2012). Reconstructing experiences with iScale. International Journal of Human-Computer Studies, 70(11), 849-865.

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J. B. (2009, April). User experience over time: an initial framework. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 729-738). ACM. doi.org/10.1145/1518701.1518814

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J. B. (2009, April). User experience over time: an initial framework. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 729-738). ACM https://dl.acm.org/citation.cfm?doid=1518701.1518814

Kim, Y. H., Kim, D. J. and Wachter, K. (2013). "A study of mobile user engagement (MoEN): Engagement motivations, perceived value, satisfaction, and continued engagement intention," Decis. Support Syst., vol. 56, no. 1, pp. 361–370. https://doi.org/10.1016/j.dss.2013.07.002

Kortum, P. and Sorber, M. (2015). Measuring the Usability of Mobile Applications for Phones and Tablets,International Journal of Human–Computer Interaction, 31:8, 518-529, DOI: https://doi.org/10.1080/10447318.2015.1064658

Krug, S. (2009) *Don't Make Me Think! A Common Sense Approach to Web Usability*, vol. Second Edi. Berkley: Newriders.

Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., & Sinnelä, A. (2011). UX Curve: A method for evaluating long-term user experience. Interacting with Computers, 23(5), 473-483 https://doi.org/10.1016/j.intcom.2011.06.005

Lewis, J. R. (2014). Usability: Lessons Learned … and Yet to Be Learned, International Journal of Human–Computer Interaction, 30:9, 663-684, DOI: https://doi.org/10.1080/10447318.2014.930311

McLellan, S., Muddimer, A. and Peres, S.C. (2012). The effect of experience on System Usability Scale ratings. *Journal of usability studies*, *7*(2), pp.56-67.

MED Device Online. (2018) *Training And Memory Decay In Simulated-Use Testing* [online]. Available at: https://goo.gl/YgdxDA (Accessed: 15 October 2018).

Moskowitz, D. S., & Young, S. N. (2006). Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry and Neuroscience*, *31*(1), 13–20.

Murphy, C (2018) *A Comprehensive Guide To User Experience Design* [online]. Available at: https://www.smashingmagazine.com/2018/02/comprehensive-guide-user-experience-design/ (Accessed: 9 April 2018).

Murre JMJ, Dros J (2015) Replication and Analysis of Ebbinghaus' Forgetting Curve. PLoS ONE 10(7): e0120644. https://doi.org/10.1371/journal.pone.0120644

Nielsen Norman Group. (2012) *Usability 101: Introduction to Usability* [online]. Available at: https://www.nngroup.com/articles/usability-101-introduction-to-usability/ (Accessed: 9 April 2018).

Nielsen, J. (2003). Usability 101: Introduction to usability. [Online] Available at: https://www.nngroup.com/articles/usability-101-introduction-to-usability/ (Accessed 26 October 2018)

Norman, D. A. (2004). Emotional design: Why we love (or hate) everyday things. New York: Basic Books.

Norman, D.A., 2009. THE WAY I SEE IT Memory is more important than actuality. Interactions, 16(2), pp.24-26.

Oishi, S., & Sullivan, H. W. (2006). The predictive value of daily vs. retrospective well-being judgments in relationship stability. Journal of Experimental Social Psychology, 42(4), 460-470.

Orfanou, K., Tselios, N. and Katsanos, C. (2015). Perceived usability evaluation of learning management systems: Empirical evaluation of the System Usability Scale. *The International Review of Research in Open and Distributed Learning*, *16*(2). DOI: http://www.irrodl.org/index.php/irrodl/article/view/1955/3262

Quote Investigator (2018) *They May Forget What You Said, But They Will Never Forget How You Made Them Feel* [online]. Available at: https://quoteinvestigator.com/2014/04/06/they-feel/#note-8611-16 (Accessed: 22 October 2018)

Sanna, L. J., & Schwarz, N. (2004). Integrating Temporal Biases: The Interplay of Focal Thoughts and Accessibility Experiences. Psychological Science, 15(7), 474–481. https://doi.org/10.1111/j.0956-7976.2004.00704.x

Sedgwick Philip. Non-response bias versus response bias BMJ 2014; 348 :g2573

Tullis, T.S. and Stetson, J.N. (2004). A comparison of questionnaires for assessing website usability. In *Usability professional association conference* (pp. 1-12).

Vermeeren, A.P., Law, E.L.C., Roto, V., Obrist, M., Hoonhout, J. and Väänänen-Vainio-Mattila, K., 2010, October. User experience evaluation methods: current state and development needs. In Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries (pp. 521-530). ACM.

Vertesi, A., Dogan, H., Stefanidis, A., Ashton, G. and Drake, W. (2018). Usability Evaluation of a Virtual Learning Environment: a University Case Study. In: 15th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA) 21-23 October 2018 Budapest, Hungary.

von Wilamowitz-Moellendorff, M., Hassenzahl, M., & Platz, A. (2006). Dynamics of user experience: How the perceived quality of mobile phones changes over time.